The 2001 NIST Evaluation Plan for Recognition of Conversational Speech over the Telephone (a.k.a. "Hub5")

– Phonetic Analysis Supplement –

Introduction

A phonetic analysis component is being supported as part of the Hub5 evaluation this year for a second time. The objective of this (non-competitive) diagnostic evaluation is to characterize ASR performance at the phonetic level. Such diagnostic information is intended to provide useful insight into model deficiencies and productive directions for improving phonetic models and overall ASR performance. Participation in the phonetic analysis component is optional. This document provides needed and useful information about the phonetic analysis component of the Hub5 evaluation. More information about the

The Tasks

The phonetic analysis component comprises a suite of three distinct and separate tasks. Participation in each of these tasks is optional and is mutually independent of the other two. In each of these tasks, phone-level output is required in addition to the usual word-level output. The tasks are:

- **Unsupervised recognition:** This task is the normal Hub5 task. The difference is simply that phone-level output is required in addition to the usual word-level output. ASR phone output is compared with reference phone transcriptions and different analyses are performed in order to assess the phonetic and articulatory variations that cause recognition errors.
- **Word-level supervision:** This task is sometimes referred to as "forced alignment". Phone recognition performance typically benefits from knowledge of the true word-level transcription. A comparison of phone recognition results here with those obtained from unsupervised recognition can help to identify deficiencies in pronunciation modeling.
- **Phone-level supervision:** This task is simply word-level decoding of speech from hand-labeled phone strings. Ideally, this should provide a lower-bound performance limit on ASR systems that represent words in terms of phone strings. A comparison of phone scores here with those obtained from word-level supervision can also help to assess the relative importance of acoustic phone models and word pronunciation models in ASR error performance.

The Data

Training and Development Data

The training and development data for the phonetic analysis component will be the same as the English data subset for the regular Hub5 evaluation. The training data comprises:

- the entire SwitchBoard (i.e., Switchboard-1) Corpus as released,
- the entire Switchboard-2 Phase-1 Corpus, and
- all English conversations of the Call_Home Corpus, including those originally designated for training and those used as test data in previous evaluations.

The development data comprises three distinct parts, namely:

- **DevSet-1:** the 20 original Switchboard-1 conversations in the 1999 EvalSet
- **DevSet-2:** the 20 Switchboard-2 Phase-2 conversations in the 1998 EvalSet
- **DevSet-3:** a set of 20 cellular conversations from the Switchboard-2 Phase-4 Corpus

Evaluation Data

The evaluation data for the phonetic analysis component will be a subset of the regular Hub5 evaluation data. Specifically, the phonetic evaluation data will comprise 7 of the 20 conversations from each of the three English EvalSets. This is in contrast to last year's phonetic evaluation, where distinct and different data were used.

Transcription Conventions

Both word-level and phone-level output will be required in order to support the phonetic analysis, and so transcription conventions are required in order to facilitate the analysis.

Word-level conventions

For ASR output, words are to be represented as defined for the regular Hub5 task. This is essentially normal orthography, as represented in the American Heritage Dictionary.[1] In addition, a special vocabulary of hesitation sounds and some speech-specific words will be accommodated. These words are listed in Table 1.

Table 1. Hesitation sounds and speech-specific words used in the reference transcription

Hesitation Words			Speech-Specific Words		
ach	eee	ew	gotta	jeeze	kinda
er	hee	hm	mhm	nah	oughta
mm	oof		wanna		

For word-level supervision, reference transcriptions will be made available. These transcriptions were produced according to two different but compatible transcription guidelines, one used at ISIP and one at LDC. These guidelines are available on the web.[2]

---

[1] A research site's vocabulary may include special nonstandard "words", including common word pairs, to optimally model particular acoustic variations. Therefore the site will need to map these variants into conventional orthography for word output and to accommodate possible pronunciation variants from conventional orthographic reference transcriptions during supervised recognition.

[2] The transcription conventions used at ISIP are described in http://www.isip.msstate.edu/projects/switchboard/doc/transcription_guidelines/transcription_guidelines.pdf

The transcription conventions used at LDC are described in http://www.ldc.upenn.edu/Projects/Cell_Trans/Transpec_cell.html

Phone-level conventions

The phonetic analysis will compare phone output from ASR systems with reference phones produced by ICSI using human phonetic transcription. The phone conventions used by ICSI are available on the web.[3] Since it is not reasonable that these conventions be adopted by all the participating sites, ICSI will provide a mapping between the phones used in the reference transcription and those submitted by the sites. In order to do this, participating sites need to submit a description of their phonetic system to ICSI in a timely fashion (according to the schedule listed in Table 2). Also, at the same time, participating sites should submit a sample phone output file to ICSI to validate file formats and phone coding.

System Output

ICSI will perform various analyses of the system output submitted by the participating sites. This system output comprises both word-level output and phone-level output. The form of the output is the same for each of the three phonetic analysis tasks, and both word-level and phone-level output are requested for each of these tasks.

Word-level output

The format for word-level output is the same as in the regular Hub5 evaluation. This is namely NIST's CTM format[4], which lists a system's best-estimate string of words to represent the speech in the source file. Each record in this CTM output file contains the following 6 whitespace-separated fields:

1. The speech waveform filename, without pathnames or extensions.
2. The waveform channel (either A or B).
3. The beginning time of the word (in seconds) measured from the beginning of the file.
4. The duration of the word (in seconds).
5. The word.
6. The confidence score for the word. This is the system's estimate of the probability that the word is recognized correctly, from 0 to 1.

Phone-level output

The format for phone-level output is the same as for word-level output. In this case, however, the file contains information about phones rather than words. Each record contains the same fields as for words, but the information is for phones. Thus the fields become:

1. The speech waveform filename, without pathnames or extensions.
2. The waveform channel (either A or B).
3. The beginning time of the phone (in seconds) measured from the beginning of the file.
4. The duration of the phone (in seconds).
5. The phone.
6. The confidence score for the phone. This is the system's estimate of the probability that the phone is recognized correctly, from 0 to 1.

[3] The ICSI phonetic transcription is described at URL: ftp://ftp.icsi.berkeley.edu/pub/speech/phoneval/STP/STP.desc

[4] The CTM format is specified in file sctk-1.2a/doc/infmts.htm in NIST's tar file: ftp://jaguar.ncsl.nist.gov/pub/sctk-1.2a.tar.Z

The confidence score

The confidence score is the system's estimate of the probability that a word (or phone) is correctly recognized, as defined in the Hub5 evaluation specification. Ideally, for the purposes of the phonetic analysis, estimation of this confidence score should be limited to sources of knowledge below the level of supervision. Thus, for example, word confidence for the word-level supervision task should not include higher level language information, and phone confidence for the phone-level supervision task should not include lexical information. This will facilitate analyses that attempt to determine the source of modeling deficiencies.

Some sites may find it impractical at this point to limit the sources of knowledge used to estimate the confidence score, or even to produce a meaningful and unbiased estimate of confidence. This is especially likely to be the case for output at the phone level. Nonetheless, it is desirable to have some indication of confidence, or likelihood, or goodness of match, so as to support an analysis of sources of error in phonetic recognition. Sites are therefore urged to provide whatever measure that they find possible and most helpful. Also, please provide a description of the measure that you choose to use.

Results submission

The phonetic analysis component will use the same submission procedures as the regular Hub5 evaluation.

Schedule

The schedule for the phonetic analysis supplement to the Hub5 evaluation is given in Table 2. This is essentially the same as that for the regular Hub5 evaluation, except that the release of supervision transcripts and the submission of results will be one week later than the regular Hub5 evaluation.

Table 2. Combined schedule for the 2001 Hub 5 regular evaluation and the phonetic analysis supplement.

Cellular DevSet Release	1 December 2000
Commitment Deadline	1 February 2001
Submission of site-specific phone set description and sample phone CTM output file	9 February 2001
Release of Hub5 EvalSet	12 February 2001
Release of phone mappings site-to-ICSI and ICSI-to-site	19 February 2001
Regular Hub5 task results due at NIST	12 March 2001 at 7:00 PM EST
Release of phonetic analysis supervision transcripts	12 March 2001
Phonetic Analysis task results due at NIST	19 March 2001 at 7:00 PM EST
Release of regular Hub5 results	26 March 2001
Release of phonetic analysis results	3 May 2001
Workshop	3-4 May 2001